CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

A Generic and Flexible Simulation-Based Analysis Tool for EMS Management

Yannick Kergosien
Valérie Bélanger
Patrick Soriano
Angel Ruiz
Michel Gendreau

December 2014

CIRRELT-2014-72

Document de travail également publié par la Faculté des sciences de l'administration de l'Université Laval, sous le numéro FSA-2014-011.

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

A Generic and Flexible Simulation-Based Analysis Tool for EMS Management

Yannick Kergosien[1], Valérie Bélanger[2,3,*], Patrick Soriano[2,3], Angel Ruiz[2,4], Michel Gendreau[2,5]

[1] Université François-Rabelais de Tours, CNRS, LI EA 6300, OC ERL CNRS 6305, 64 av. Jean Portalis, Tours, 37200, France

[2] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

[3] Department of Management Sciences, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, Canada H3T 2A7

[4] Department of Operations and Decision Systems, 2325 de la Terrasse, Université Laval, Québec, Canada G1V 0A6

[5] Department of Mathematics and Industrial Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

Abstract. Emergency medical services (EMS) are dedicated to provide urgent medical care to any person requiring it and to ensure their transport to a hospital or care facility, if required. Moreover, in many contexts, EMS also have to provide transportation services for patients needing to go from one hospital to another or between their home and the hospital. For such organizations, efficient strategies for managing the ambulance fleet at their disposal have to be selected, but the highly random and dynamic nature of the system under study makes this a challenging task. Most of the published studies which have considered these issues have done it focusing on a specific EMS context, one city or one territory for instance. However, it is possible to identify several common characteristics and processes from one EMS context to another. This is the purpose of the generic discrete event simulation-based analysis tool proposed here, which can be adapted to a wide range of EMS contexts. In particular, it explicitly considers the two types of tasks that can compose the mission of an EMS: serving emergency requests and providing transports between care units/hospitals/patients' homes.

Keywords. Emergency medical services (EMS), simulation, ambulance fleet, management strategies.

Acknowledgements. This research was supported by the Fonds de recherche du Québec – Nature et technologies (FRQNT) under grant PR-122269. This support is hereby gratefully acknowledged. The authors also wish to thank the anonymous referees for their valuable comments.

---

* Corresponding author: Valerie.Belanger@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2014

1 Introduction

An important entry point and a critical element of modern health systems is the pre-hospital part which is commonly known in North of America as Emergency Medical Services (EMS). The main objective of an EMS is to provide basic medical care for any person requiring it at the site of an emergency and to transport these patients to a hospital or care unit, if needed. Furthermore, for critical cases, the response time to provide first care is a crucial element that can greatly influence the patient's health and recovery. In order to ensure an adequate service level to the population of the region they serve, such organizations have to mobilize and manage efficiently considerable resources (*i.e.* paramedics, ambulances, emergency medical responders, *etc.*). Doing so is an extremely complex task due to, among other reasons, the uncertain nature of the emergency calls concerning both the arrival time as well as their locations.

In addition to this first task, many EMS are also in charge of a second type of service, which consists in transporting patients between different care units/hospitals or eventually to or from their homes. These transportation demands or transfer demands, as they will be referred to here onwards, are generally received dynamically, but sufficiently in advance so they can be scheduled, which is not the case for emergency demands. In several EMS contexts, the same fleet of vehicles is used to carry out both types of tasks giving rise to complex decisions regarding fleet management strategies. For simplicity reasons, most EMS just split the overall fleet into two subfleets, one assigned to emergency calls and the other to transfer demands. Each fleet is then managed independently. Alternative fleet management strategies such as partial or complete pooling of ambulances can also be considered. These strategies could eventually result in a more efficient management or a reduction in the fleet size required to achieve a given service level. However, to evaluate such alternatives, one needs powerful analysis tools. Numerous problem solving tools have been proposed so far, but simulation is one of the most widely used approaches (Aboueljinane et al., 2013). Simulation allows to easily integrate stochastic and dynamic aspects faced in EMS environments. However, most of the simulation studies conducted to address problems arising in these cases are generally concerned with the specificity of a given city or territory (Mason, 2013). Even if each EMS operates in a particular and somewhat different context under its own management rules, it is generally possible to identify several common characteristics and processes from one EMS context to another. Hence, it would be very useful to develop a simulation-based analysis tool that is generic and flexible enough to easily adapt to many EMS contexts.

The main objective of this paper is therefore to propose a generic discrete event simulation-based analysis tool that can be adapted to a wide range of EMS contexts. One interesting aspect of the proposed simulation model is that it explicitly considers the two possible tasks that make up the mission of an EMS: serving emergency requests and providing transports between care units/hospitals/patients' homes. In previous EMS simulation models, these two types of tasks have generally been addressed separately. The proposed simulation model is therefore highly flexible allowing the analysis of several management strategies for both types of requests, either considered together or independently. Moreover it can complement another solution tool in a simulation-optimization scheme to validate the results obtained while considering the dynamic and stochastic aspects inherent to EMS contexts.

The paper is organized as follows. Section 2 provides an overview of previous related simulation studies published in the literature. Section 3 outlines the main EMS characteristics such as its environment, actors, processes, and decisions. The simulation model is then fully described in section 4. Finally, in order to verify and validate the simulation model and to show its capabilities, section 5 presents a series of computational experiments. Concluding remarks and future research avenues are presented in the last section.

2 Literature review

The literature related to EMS is vast and many studies continue to appear regularly. Most studies related to EMS are specifically focused on the ambulance location and the ambulance relocation problems. The ambulance location problem consists in selecting the potential standby sites as well as determining

how many ambulances should be located at each of them in order to ensure an adequate coverage of the population. Once implemented, this location plan remains unchanged. The ambulance relocation problem concerns the relocation of ambulances to standby sites in order to consider the evolution of the system over a day. Such relocations aim to maintain an adequate service level at all times. Other problems, such as the dispatching problem which consists in selecting the right ambulance to dispatch to the scene of an emergency, and the fleet management decision problem which consists in selecting which type of task to assign to each ambulance, arise in the EMS context. However, studies that explicitly analyze those types of decisions are rather scarce compared to the ones dealing with location decisions.

Literature reviews have been published over the past years focusing on both ambulance location and relocation problems from different methodological standpoints. ReVelle et al. (1989), Marianov and ReVelle (1995) and Brotcorne et al. (2003) present an interesting overview of mathematical models applied to ambulance location problems. Goldberg (2004) and Bélanger et al. (2012) propose a review of the different approaches developed to tackle location and relocation problems with most of these approaches falling within the field of mathematical programming, queuing theory and simulation. Finally, Aboueljinane et al. (2013) propose a complete survey of simulation models applied to EMS operations. In this section, we will provide a brief review of studies that consider simulation techniques to address various problems related to EMS management. This review is not meant to be exhaustive. It rather focuses on the most relevant works related to the purpose of this paper. We refer those who are interested in a more detailed description of mathematical models to the reviews cited above.

As stated earlier, simulation is a widely used approach in the field of EMS, either to test different alternatives or to validate/evaluate solutions obtained by solving mathematical models. One of the first attempts to address location decisions in an EMS context through the use of simulation is due to Savas (1969). The aim of the study was to evaluate the possibility of introducing a second site where ambulances could wait. Since then, many researchers have used simulation to analyze and validate different decisions faced by EMS such as the selection of location plans (Swoveland et al., 1973; Berlin and Liebman, 1989; Lubicz and Mielczarek, 1987; Fujiwara et al., 1987; Goldberg et al., 1990; Harewood et al., 2002), the dimensioning of the ambulance fleet (Liu and Lee, 1988) and the selection of management strategies (*e.g.* dispatching rules (Gendreau et al., 2001; Carpentier, 2006; Andersson, Petersson, and Värbrand, 2007) and relocation strategies (Repede and Bernardo, 2008; Carpentier, 2006; Rajagopalan et al., 2008; Gendreau, et al., 2006; Andersson, Petersson, and Värbrand, 2007)). Some authors also propose simulation-based analysis tools that are able to consider together or independently several types of decisions (Trudeau et al., 1989; Goldberg et al., 1990; Ingolfsson et al., 2003). More recently, Henderson and Mason (2005) present a decision support tool developed for the ambulance service of St John (New Zealand) that benefits from the combination of simulation and specialized data visualization tools (GIS). The flexibility of this tool also seems to be one important asset since it allows its direct application to other cases as presented in Mason (2013). Finally, Zhen et al. (2014) propose a simulation model embedded within a simulation-optimization framework in order to determine the best possible ambulance deployment in a stochastic environment. Table 1 summarizes the main decisions, system characteristics and assumptions considered in the development of the simulation model proposed in these studies, and helps position our work with respect to the existing literature.

As can be observed, most studies on ambulance fleet management have focused exclusively on emergency demands. Few studies dealing with transfer demands have been addressed in healthcare contexts have been studied so far (Beaudry et al., 2009; Hanne et al., 2009; Parragh, 2011; Kergosien et al., 2011) and they are generally treated as variants of dial-a-ride problems. Both types of problems have mostly been studied independently. However, alternative fleet management strategies such as partial or complete pooling of ambulances can also be envisonned. These strategies could eventually result in a more efficient management or a reduction in the fleet size required to achieve a given service level, but at the price of more complex managerial and decisional processes. Unfortunately, we found very few tools or research explicitly addressing this issue and the underlying trade-offs between centralization and decentralization of resources in the context of EMS. To the best of our knowledge, the only paper that deals simultaneously with both types of demands is Kiechle et al. (2008). In that study, the authors test different strategies

	Savas	Swoveland *et al.*	Berlin and Liebman	Lubicz and Mielczarek	Fujiwara *et al.*	Liu and Lee	Trudeau *et al.*	Goldberg *et al.*	Repede and Bernardo	Gendreau *et al.*	Harewood *et al.*	Ingolfsson *et al.*	Henderson and Mason	Carpentier	Andersson *et al.*	Rajagopalan *et al.*	Mason	Zhen *et al.*	Kergosien *et al.*
Type of decisions or analysis																			
Fleet dimensioning	■					■	■				■	■	■				■		■
Location decisions	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■
Relocation strategies							■		■	■			■	■	■		■		■
Dispatching rules		■					■			■			■	■	■		■		■
System characteristics																			
Districting		■		■				■			■				■		■		
Priority of calls		■		■			■		■	■		■	■		■		■		
Type of requests																			
Emergency	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Transfer				■									■				■		■
Fleet management strategy (if transfer requests considered)																			
Independent													■						■
Transfer as emergency				■															■
Other																			
Location of ambulances																			
Static location	■	■	■	■	■	■	■	■			■	■	■	■	■	■	■	■	■
Relocation strategies						■		■	■			■	■	■	■		■		■
Dispatching rules																			
Nearest ambulance	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■
Other									■					■	■		■		■
Input data																			
Demand arrival (inter-arrival times)																			
Historical data							■						■			■	■		
Deterministic	■																		
Empirical dist.								■											
Poisson process (exp. dist)		■	■	■	■	■			■	■	■	■		■				■	■
Unspecified															■				
Travel times																			
Fixed matrix		■													■				
Historical data																	■	■	
Weibull dist.			■																
Gamma dist.									■										
Linear regression								■											
Distance and speed*				■	■		■			■	■	■						■	
Complex computation													■				■		
Modified euclidean																			■
Unspecified	■					■									■		■		
Intervention times												■							
Deterministic																			
Empirical dist.								■											
Uniform dist.						■													
Exponential dist.			■							■									
Normal dist.			■																
Gamma dist.									■										■
Historical data																■			
Unspecified	■	■		■	■		■				■		■	■	■		■	■	

Table 1: Summary of simulation studies

*Methodology used to compute distance and determine speed vary.

of ambulance movement based on the selection of standby points that can only be hospitals. However, no step for solving a location problem is considered.

We can conclude that there is still a need for generic simulation models applicable to a wide spectrum of contexts or that, at least, can be easily adapted to them. This is the aim of the present study which, in particular, is able to consider the two types of tasks performed by EMSs, *i.e.* emergency call response and transfer demands, enabling future studies on a whole array of different management strategies, ranging from independently managed fleets to fully integrated ones.

3 EMS components

This section presents the main components related to the management of an EMS: the EMS environment and actors, the emergency and transfer demands processes as well as the various decisions to be taken. These three components and their relations allow characterizing an EMS.

3.1 EMS environment and actors

An EMS and its environment can usually be summarized by the following elements:

- **Zones**: Zones are the basic subdivisions of the region to be covered. To each zone is associated a specific territory often described by its centroid as well as the population to serve within it. The notion of coverage is defined with respect to these zones: a zone is said to be fully covered if and only if an ambulance can reach all demands originating in that zone within the prescribed delay.
- **Districts**: Most EMS also divide the region they need to cover into several districts, each district being composed of one or several zones.
- **Hospitals**: Hospitals are the facilities that have the care units needed to receive patients following an emergency demand. They are also points of departure and/or arrival of patient transfers.
- **Potential standby sites**: Sites located at strategic locations in the region covered by the EMS where one or more ambulances can park while waiting for emergency calls. Hospitals are obviously potential standby sites.
- **Depots**: A depot represents a location where ambulances start and finish their shifts. There may be several depots in the region covered by the EMS.

The main actors involved in an EMS are:

- **Emergency medical responders** (EMR): EMR are the persons who respond to emergency calls. Their role is to draw up a health check, provide help and advice by phone, determine the priority of the call depending on the patient's condition, and decide if an ambulance must be sent to the site where the patient is located or not. If this is the case, then the call is forwarded to an operator.
- **Operators**: Operators have the main responsibility of selecting the ambulance to be dispatched to the scene of the emergency once an emergency call has been transferred by an EMR. Computer aided dispatch systems (CADS) are sometimes available to guide the operators to take the best possible decisions. CADS usually display the location of the demands to be processed and propose a list of ambulances that can be dispatched in order to adequately serve the demand. Operators also have to manage the ambulance fleet and thus take decisions according to the different ambulance movements over time such as real-time relocation decisions. If the region to cover is divided into several districts, an operator can be in charge of one or several districts.
- **Paramedics and ambulances**: Ambulances are the vehicles used to transport patients. They carry the medical supplies and equipment required for on-site treatment and they are able to transport one patient at a time. Each ambulance is manned by a crew generally composed of

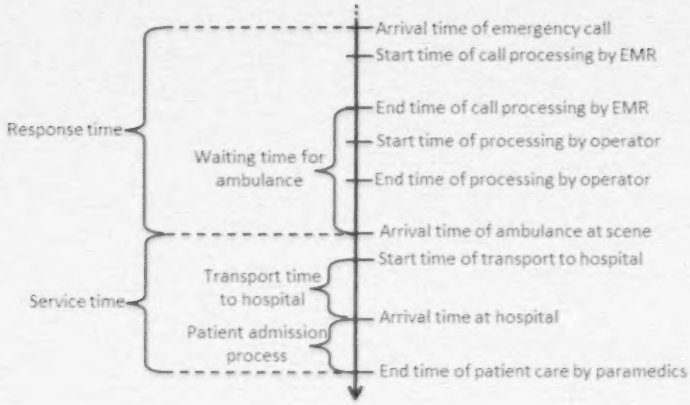


Figure 1: Emergency demand process stages

two paramedics qualified to provide advanced medical care for emergency cases. The ambulance fleet can either be homogeneous or heterogeneous. In the latter case, several types of ambulances having different characteristics (e.g. medical equipment, size, etc.) are available. Depending on the context, the number of available ambulances can vary over time depending on the work schedules of paramedics. If the region to cover is divided into several districts, an ambulance may be assigned to one or several districts.

The three main types of actors just described are defined from a functional point of view. Depending on the EMS context under study, one can find organizations where the processes may differ from the ones described herein, but they will generally contain these three functional actors.

3.2 Emergency and transfer demands processes

Demands treated by an EMS can be classified into emergency demands and transfer demands.

3.2.1 Emergency demands

Emergency demands arriving to the EMS call center can either require medical advice or the assistance of an ambulance at the site of the emergency. In the first case, the demand process ends after it has been processed by the EMR. In the second case, the EMR determines the priority of the call and transfers it to an operator who will select the proper ambulance to dispatch to the call. Once the ambulance arrives on the scene, the paramedics will provide the required medical treatment to the patient. If no transport is required, the demand process ends. Otherwise the ambulance will transport the patient to a specific hospital, determined according to several criteria such as distance, patient pathology, *etc.* The hospital destination can be chosen by paramedics, operators or EMRs. It can be consider as a data or a decision to be made. After arriving at the hospital, the patient is transferred to the staff of the receiving care unit. Figure 1 summarizes all the stages involved in the emergency demand process and their duration. It also introduces the response and service times that constitute the most commonly used operational performance measures. The response time is generally considered the most important criterion to assess EMS effectiveness. The arrival time of the calls and the time spent at each stage are not known in advance.

3.2.2 Transfer demands

Transfer demands are usually received dynamically by the emergency medical responder or directly by the operators. When the demand is received, an operator will determine which ambulance will perform the transportation request when needed. The type of ambulance required, the priority, the starting and

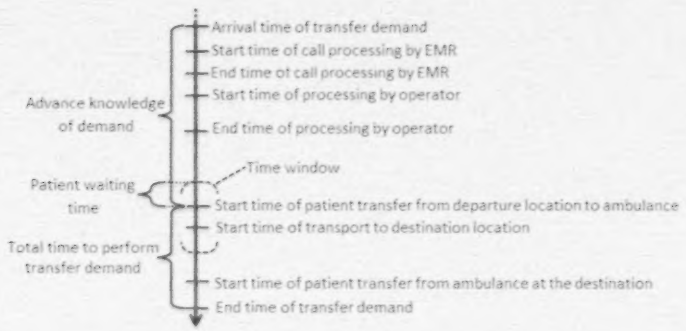


Figure 2: Transfer demand process stages

destination points, the time window during which the transport has to begin in order to be on time at the destination point, the ambulance availabilities, and the demands already planned will be considered in the dispatching process. Once the time at which to perform the demand has been reached and the selected ambulance is available at the departure location, the transportation takes place. This transport is divided into three main steps. The first part consists in the transfer of the patient from its departure location to the ambulance. The second corresponds to the ambulance travel to the destination location. Once the ambulance arrives at the destination, the last part concerns the transfer of the patient to its service destination (usually by stretcher or wheelchair) (cf. Figure 2). However, a transfer demand may be canceled at any time before the patient is loaded into the ambulance, forcing the operators to revise the planning.

3.3 Decision making

To ensure an adequate service level to the population, an EMS has to mobilize several resources (*i.e.* paramedics, ambulances, EMRs, operators, etc.) and then, manage them efficiently. Therefore, several questions arise regarding the means and strategies to be deployed in order to respond efficiently to the demands received, for instance: How much of each resource should be mobilized? How should the ambulance fleet be managed? Where should the ambulances be located? All such decisions can be classified according to three levels of decision-making:

- **Strategic level**: Long-term decisions address the location of the depots or call centers, the type of management, the dimensioning of the ambulance fleet (for each type of vehicle), the determination of the staffing levels and the division of the territory into districts.
- **Tactical level**: Medium-term decisions involve the location of the potential standby sites, the selection of staff management strategies (crew pairing and scheduling), and the allocation of the ambulances to task types (emergency or transfer demands) as well as their eventual allocation to potential standby sites.
- **Operational level**: These short-term decisions concern the management rules such as dispatching decisions, choice of hospital, redeployment policies, break scheduling, and the scheduling of transfer demands.

The decision-making process is very important in the EMS context. Indeed, the choices and decisions made will directly impact the quality of service for both types of demands as well as the operation costs. Moreover, these decisions are usually closely related and may have significant impact on each other. In order to better assess and analyze the impact of each decision on the performance of the system, we propose the development and use of a generic simulation model.

The simulation model proposed in this work can be used to consider decisions at the three levels either by changing the input data or decision routines. At strategic and tactical levels, input data such as the

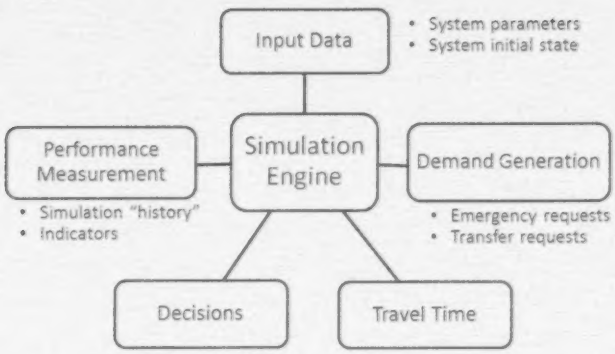


Figure 3: Simulation model architecture

number of ambulances or the allocation of ambulances to task types can be modified to address different situations. At the operational level, decision routines that replicate the policies or strategies to evaluate can be adapted to the context under study. The simulation model thus aims to be flexible enough to address decisions taken at all levels, but also to evaluate the interaction between decisions at the different levels. In particular, and in order to illustrate the potential of the proposed simulation model, Section 5 evaluates two strategies to respond to urgent and transfer requests.

4 Simulation model

An important feature of our simulation model is that it separates 'physical' parts (the resources) from 'decisional' parts: doing so allows us to better understand EMS management strategies and also increases the flexibility and generality of the model. Taking this into account, the proposed EMS object-oriented model is built on the following components: an *Input data* block containing all the system parameters as well as the information on the system state at the beginning of the simulation, a *Demand generation* block that provides the emergency (i.e. urgent) and transfer requests, a *Simulation engine* which manages the simulation clock, a *Travel time* block which estimates ambulance travel times between sites, a *Decisions* block which decides, according to the current fleet situation, the different tasks and activities to be assigned to each particular ambulance, and finally, a *Performance measurement* block which traces and compiles all the information required to evaluate the system performance. The model architecture is illustrated in Figure 3.

4.1 Input data

Input data can be classified into two groups. The first one defines the system and its configuration, in particular the division of the region under the EMS responsibility into zones and districts, the list of available ambulances, the set of potential standby sites, and the set of hospitals in the region. A zone is characterized by its geographical Cartesian coordinates, a population density, and a probability vector. The probability vector of a given zone x represents, for each period, the probability that the next request is located in x. Since this probability can evolve according to the time of day, a day is decomposed into several periods whose length can be fixed by the user. To each zone is also associated a set of potential standby sites from which the zone can be reached (covered) within some pre-specified time delay (the covering time), or several of such sets if more than one covering time is used. The set of standby sites contains their geographic location as well as the maximum number of ambulances that can be stationed at each site. Each hospital is characterized by its location and a probability vector containing, for each time period, the probability that a transfer demand originates at that hospital. The second class of data defines the initial state of the system. It includes the initial position of ambulances, the initial state of each resource, etc.

4.2 Demand and random variable generation

Simulation models uncertain events by means of probability distributions. An important part of simulating consists in sampling these probability distributions in order to draw random variable realizations or, in other words, plausible specific values for uncertain events to be used during the simulation execution.

For example, service time at a patient site is, in practice, uncertain. We model service times by a probability distribution function of known parameters. When we simulate ambulance activities, we sample this function to generate the actual time that the ambulance will spend at the patient site. Note that several random values are set in advance, before the simulation starts, but will only be used by the system later on. For instance, although a request may not require the transport of the patient to a hospital, an ambulance needs to be sent to the site and only after the paramedic team has arrived will it become known whether or not the transport is required.

The reason why some random events and values, such as service times, are not generated dynamically during the simulation but drawn a priori and stored as input data files is twofold. First, this approach is very useful during the model development because it allows to validate/verify the model behavior and to debug the code if required. Second and more importantly, generating as much random events as possible out of the simulation execution drastically reduces the variance of the simulation results. Indeed, when comparing the performance of two management strategies or configurations, one never knows which part of the measured differences is due to the particular values taken by random variates in each simulation and which is due to the differences between the tested alternatives. Variance reduction techniques promote the use of independent random numbers between replications of a same experiment, but of strongly positively correlated random numbers between runs of different alternatives. Using the same variates for all the configurations under study clearly minimizes variance of the results.

4.3 Simulation engine

The model proposed in this paper is based on discrete event simulation (DES). DES, as defined in Law and Kelton (2000), deals with the modeling of a system as it evolves over time by a representation in which the state variables change instantaneously at particular points in time and where the system is defined by a set of entities each characterized by a set of attributes and a set of state variables. These points in time are the ones at which an event occurs, where an event is defined as an instantaneous occurrence that may change the state of the system. The events management is provided by a *simulation engine*, a timing routine which is in charge of moving the simulation clock from one event time to the following. Each time an event occurs, some decision procedures are triggered and the state of the system and it's entities are modified or adjusted according to the decisions taken. The simulation process then resumes moving the simulation clock to the next event.

Model entities. From an implementation perspective, the simulation model is composed of programming 'blocks' which represent entities and servers within an EMS. Entities have fixed attributes (i.e. characteristics that do not change during the simulation execution) and states (i.e. characteristics which evolve during the simulation execution). The model uses two main entities, *Requests* and *Ambulances*.

Request entities can take three states only: waiting, in treatment and done. However, they are characterized by a large number of attributes, among which it is worth to mention:

- Type of request : emergency or transfer,
- Diag : hospital or abort (indicates if the patient of an Emergency request needs to be transferred to the hospital, and if yes, to which one),
- Arrival time, Treatment time at scene (for emergency requests).

Since we use an Object-Oriented approach, the same "request" structure is used to represent an unlimited number of requests which differ one from the other by the specific values taken by their attributes

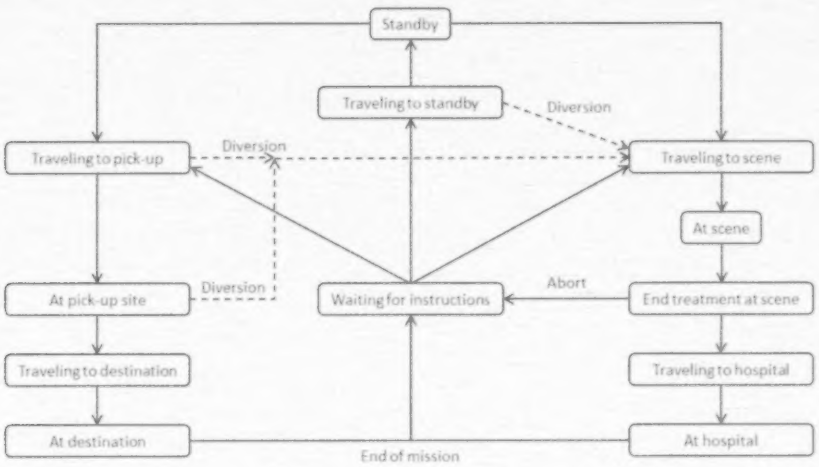


Figure 4: Possible states for an ambulance

(origin zone, call time, request type).

Ambulance is the second type of entity. A generic Ambulance also has specific attributes (for example, it is assigned to a fixed depot and works under a given fixed schedule) and may evolve through different states (it may be idle, in transit towards a standby point, or even responding to an emergency request) as the simulation runs. More precisely, Figure 4 illustrates the possible states of an ambulance. The left part of the figure describes the cycle of states associated to a *transfer* request, while the right part depicts the one for an *emergency* request. Transitions between states follow from the realization of the events.

Discrete-event simulation engine. The simulation engine described in this paper is inspired from the one proposed by Pidd (2004), which consists of a three-phase algorithm that allows the clock to be advanced asynchronously from one event to the next. The simulation engine works with a list of events sorted by their execution time. Each time an event is executed, the system state (and eventually the entities states) is modified accordingly, and the simulation clock moves to the following event in the list. Sometimes, the execution of an event does not change the state of the system, but rather generates new events to be added to the list. In this case, the time at which the event will be executed is computed or simulated (drawn from an appropriate probability distribution). Events are classified into bounded (B) and conditional (C). B-type events are those for which the execution date can be predicted (or simulated) by the system. For example, let t be the current time at which a particular event (e.g., *departure from standby site towards request site*) is executed. The execution of this event implies changing the ambulance state from *idle* to *traveling to scene*, changing the request state to *in treatment*, and the generation of a new event (e.g., *arrival to request site*) whose execution date can be stated as $t + d$, where d is the travel time between the ambulance standby site and the location associated with the specific request. Regardless of the approach followed to model travel times, which can be either deterministic or stochastic, d is a computed or *simulated* value. On the other hand, the execution date for conditional events cannot be determined in advance because they depend on the current system state. For example, it is not possible to set the execution date of an event like *departure from standby site towards request site* in advance because the event can only be executed when the following two conditions are satisfied simultaneously: (1) an emergency request is waiting to be served, and (2) an ambulance able to serve that request is available. By defining different condition sets for conditional events, it is quite straightforward to integrate specific behaviors or management strategies.

Table 2 illustrates the possible states for the main system entities (request and ambulance). We noted

E an emergency request, T a transfer request, D a dispatcher (operator) and A an ambulance.

4.4 Decisions block

The execution of some events implies modeling and reproducing some of the decision processes carried out by operators or dispatchers. Indeed, these decision processes mimic the practices and tactics followed by the EMS organization under study. In fact, one might see the Decisions block as the "expert" to whom the simulator turns to when choices need to be made like, for example, selecting the ambulance to respond to an urgent request. This "expert" may take the form of a mathematical program, an artificial intelligent algorithm, or in the simplest case, decisions rules. These decisions, as well as the specific choices selected in the context of this study will be discussed in Section 5.3.

4.5 Travel time calculator

Simulating travel times is a very difficult yet important task. In fact, travel times influence the precision of the simulation results and are a key element when evaluating the credibility of the simulator. Although most of the simulators in the literature use pre-computed travel times between pairs of pre-determined points (zone centroids, standby points, and hospitals), in our context, the fact that an ambulance may be diverted while it travels towards a new destination, implies that one needs to be able to evaluate where the ambulance is at the moment it is diverted in order to compute a precise travel time from that point to the new destination. Several methods can be used to estimate travel times, including sophisticated methods linked to powerful geographic information systems (GIS). We elected to implement the following relatively simple and generic method based on *a priori* knowledge of some real travel times or estimates for a set of important or frequent locations (e.g. hospitals, potential waiting sites, zone centroids, etc.). Let M be the matrix of known travel times between these locations, the size of M depends on the amount of information that can be obtained from the real case studied. Evidently, increasing the number of points in M will increase the accuracy of the estimated travel times. During the simulation, the computation of the travel time t_{ab} between two locations a and b not in matrix M, is based on the known travel time $t_{a'b'}$ between two locations a' and b', where a' and b' are the locations in M that are the nearest to a and b respectively, and on the Euclidean distances from a to b and from a' to b' noted d_{ab} and $d_{a'b'}$, as follows :

$$t_{ab} = \frac{d_{ab} * t_{a'b'}}{d_{a'b'}}$$

Clearly, this method is not as accurate as a GIS based one. However, if matrix M contains enough points, this method should approximate adequately travel times by taking into account through the data in the matrix the presence of obstacles or particular features of the transportation infrastructure (e.g. highways, bridges, tunnels, one-ways etc.) as well as the general traffic conditions on the itineraries corresponding to each pair of locations in M. When an ambulance is diverted, its current position is estimated at a distance equal to α from the original point in direction of the destination point, where α is the ratio between the elapsed time since the ambulance left its departure point and the total travel time to the destination point.

4.6 Performance measurement

The last block is devoted to support the analyses of the simulation results by the user. To this end, a complete history of the simulation is used to calculate some performance indicators. This history includes all movements and demands performed by each ambulance, all the times at which the entities and resources states changed, and other statistical information about the decisions taken, e.g. the number of deployments. Some of the performance indicators offered to the user are: the response times to answer calls by EMRs and operators, the elapsed time between the arrival of an emergency call and the arrival of the ambulance at the accident scene for emergency demands, the delays for transfer demands, the workload of each team and amount of overtime if there was any, the number of times an ambulance was redeployed or diverted. The list of movements of each ambulance can in particular be used to visualize the work of ambulances through a graphical interface.

Event	Type	Condition	Decisional procedures	States changes	Events
Dispatch E (or T)	C	E (or T) = waiting & D = available		E (or T) = in treatment; D = busy	Assign ambulance to E (or T)
Assign ambulance to E	B		Scheduling and Find A	A = traveling to scene; D = available	Arrival at scene
Arrival at scene	B			A = at scene	Treatment at scene
Treatment at scene	B			A = end treatment at scene	
Departure for hospital	C	A = end treatment at scene & diag = hospital		A = traveling to hospital	Arrival at hospital
Abort request	C	A = end treatment at scene & diag = Abort		A = wait for instructions; E = done	End of request E
Arrival at hospital	B			A = at hospital; E = done	End of request E
Assign ambulance to T	B		Scheduling and Find A	A = traveling to pick-up site; D = available	Arrival at pick-up site
Arrival at pick-up site	B			A = at pick-up site	
Departure for destination	C	Patient = ready & A = at pick-up site		A = traveling to destination	Arrival at destination
Arrival at destination	B			A = wait for instructions; T = done	End of request T
End of request E	B		Assign standby site	A = traveling to standby	Arrival at standby
Arrival at standby	B			A = standby	
End of request T	B		Assign transfer request or standby site	A = traveling to pick-up site or traveling to standby	Arrival at pick-up site or at standby
Perform diversion	C	Diversion = 1 & (A = at pick-up site, A = traveling to pick-up site, A = traveling to standby)		A = traveling to scene or another pick-up site	Arrival to scene or at pick-up site
Perform relocation	C	A = standby	Relocate ambulance	A = traveling to standby	Arrival at standby

Table 2: Main events

The next section proposes some computational experiments which will allow to illustrate how the generic simulator may adequately reproduce any real or pseudo-real environment. Before presenting the computational results, we will discuss how the model was verified and validated.

5 Computational experiments

This section presents several numerical experiments illustrating the usefulness and flexibility of the simulation tool. However, it is not meant to show wether a given management policy is better than another. For this, the simulator would need to be fed with the real data and decision rules of a particular organization which is not the case here. The section describes also how verification and validation of the simulation model were conducted.

5.1 Implementation, verification and validation

The simulation model was implemented in C++. The use of a generic programming language was justified by the need of higher flexibility and to avoid the restrictions due to specific architectures of simulation softwares. Moreover, this approach allows us to easily build routines that will replicate almost any decision procedure.

The verification of the simulation model consists in the following activities: inspecting simulation program logic, performing simulation test runs and inspecting sample path trajectories, and performing simple consistency checks (Altiok and Melamed, 2007). In order to ease the verification process, the simulation tool has been designed to store the complete history of all the events and decisions performed during the simulation experiment, *i.e.* the sequence of tasks and movements performed by each ambulance as well as the information regarding each specific demand. Thus, the sequence of events for some specific ambulances or demands can be traced to make sure that the implementation is correct and the simulation works the way it is supposed. Several functions have been implemented to check the correctness of the simulation process. Among others, functions check if the succession of entity/resource states are correct, if the ambulance routes are feasible with respect to times and locations, if all demands are performed as they should (e.g. on the right day, at a reasonable time, with emergencies requiring transport to a hospital effectively followed by such a transport, *etc.*).

5.2 Test data

We decided to base our test data on a real city in order to adequately represent EMS issues such as high and low density zones, the presence of a downtown sector, and so on. Therefore, the context of Montreal and Laval (the suburb just north of Montreal), which constitute the major population center in the province of Québec (Canada) with about 2.3 million inhabitants (Urgences-Santé, Rapports annuels), was elected. However, the EMS described here is fictitious in the sense that the strategies and rules that implement the operational and real-time management of the system such as fleet management strategies, location and relocation policies, as well as dispatching decisions, are based on a set of rules generally considered and accepted in the literature rather than the ones actually used by the local EMS of Montreal (Urgences-santé) for which we have no official information. Thus, two sub-fleet of ambulances (one for emergency and one for transfer) were considered, the nearest ambulance is always dispatched to a call and the location of ambulances are determined in real-time based on the system state, which seems to be in line with the strategies used by the local EMS. Nevertheless, the way each of these decisions are effectively taken is based on models proposed in the literature which will be discussed in the next subsection.

Figure 5 shows the region covered by the fictive EMS where each dot represents the gravity center of a zone and the size of a dot indicates the relative importance of the population in that zone. The region contains: 600 zones, 40 potential standby sites arbitrarily located, two depots, and 15 hospitals.

In the experiments, we consider that the fleet is managed in a "centralized" manner, meaning that a sole decision maker manages the whole region and all the ambulances. However, the simulation can

Figure 5: Cartography of the region to cover

easily be parameterized to consider several districts. Also, since no real data was available to us regarding the demand, ambulances travel times or service times, we randomly generated a set of realistic data by merging several sources of information: annual reports of the local EMS organization Urgences-Santé (Rapports annuels), population statistics for the region (Statistics Canada (2012)), and information collected from the literature. Most of these informations being of an aggregated nature, we therefore set the parameters of our generator empirically in order to ensure that the detailed data generated was effectively realistic and adequately fitted the aggregate data collected (in terms of total number of requests of both types, number of transports, number of teams and size of the fleet).

As is generally accepted in the literature (Ingolfsson, 2013), an exponential distribution was used to model the inter-arrival times between two consecutive emergency demands. We divided each day into 12 two hour periods to build a daily workload curve which accounts for the variation in demand intensity throughout the day. The mean of the exponential distribution was arbitrarily set to a specific value for each of these periods, ranging from 1.5 to 5 minutes. As suggested in Law and Kelton (2000), we used the method described in Lewis and Shedler (1979) to generate arrivals from the resulting non-stationary arrival process. Once a request is generated, we use the process depicted in Figure 6 to set its type, attributes, and sampled probability distributions. For example, once the arrival of a new request is generated, its type is set to *Urgent* with a probability 0.75 or *Transfer* with probability 0.25. If the request is set to Urgent, then we randomly decide if the request will need transportation to a hospital or not (probabilities of 0.75 and 0.25, respectively). If a Transport to a hospital is generated, then *Time at scene* and *Discharge time at hospital* are drawn from Gamma distributions according to the observations in Schmid (2012).

An urgent request is associated to a specific zone following a discrete distribution where the probability of selecting a zone depends on its demographic weight.As discussed in (Aboueljinane et al., 2013), this is one of the approaches proposed to adequately generate demands. For transfer requests, their origin location (or destination) is randomly determined to be a hospital with probability 0.85 (the specific hospital also being selected randomly) or a patient's home (with probability 0.15). In the latter case, the specific coordinates of the patient's home are generated uniformly.

We also assumed that call handling by EMRs and operators is not a bottleneck, so the number of EMRs and operators in the simulation model were set to values large enough so that incoming calls

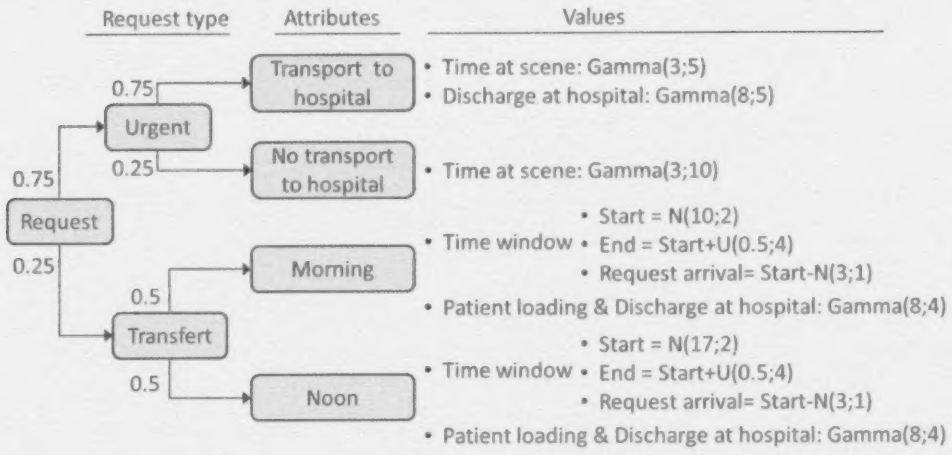


Figure 6: Generation process of attributes and random values for a request

would generally not have to wait before being answered. Call processing times by EMRs and operators are generated according to a Gamma(1;2) distribution.

Travel times are computed as described at the end of section 4.5 using the Euclidean distance between each point. It is important to recall that the demand generator as well as the travel time computations are independent from the simulation tool itself. Different probability distributions or parameters can easily be used to model other contexts. Moreover, GIS based methods could also be used to compute more accurate travel times.

5.3 Decision procedures

Within each simulation run, several decisions need to be made to adequately handle both types of demands. This section discusses the strategies adopted to tackle these decisions as well as the algorithms used to solve the underlying problems.

Our experiments will compare two different fleet management strategies. The first strategy considers that ambulances are separated into two fleets, one assigned to emergency demands and the other assigned to transfer demands, and that the two are managed independently. The second strategy considers a complete pooling of ambulances where both emergency and transfer demands can be assigned to any ambulance. Let us describe first the case of independent fleets.

Decision procedures of the independent fleet management: In the case of emergency demands, the decisions that have to be made are vehicle deployment, redeployment and dynamic redeployment eventually, as well as deciding which vehicle to assign to each request. In fact, each time an ambulance becomes available (i.e. starts its working shift) or unavailable (i.e. it is assigned to a mission or it ends its working shift), or whenever the coverage has degraded under a given threshold, the system considers where to deploy or redeploy the whole fleet. The relocation problem is based on the one proposed in Gendreau et al. (1997), in which two types of covering constraints are considered, as recommended in the United States EMS act (United States EMS Act, 1973): absolute covering constraints requiring all demands to be reachable by at least one ambulance within a given time limit r_2 and relative covering constraints stating that a proportion α of all demands have to be reachable within a given time limit r_1, with $r_2 > r_1$. In our case, r_1 is set to 9 minutes, r_2 to 11 minutes and α to 90% of the demands. The model aims at maximizing the sum of the zones that are covered twice within r_1 minutes weighted by

the probability of a new demand occurring in that zone. This performance measure is often used in the literature to take into account the stochastic nature of ambulance availability (i.e. although located to cover a given zone an ambulance might eventually be unavailable if it is already answering a previous emergency). By using the maximum double coverage one seeks to maintain coverage of demand areas as high as possible even though some ambulances may be already responding to calls. This idea was first introduced by Hogan and ReVelle (1986) and used in several studies among which the one of Gendreau et al. (1997) that was selected to illustrate the simulation-based analysis tool. The relocation plan is computed and applied if the coverage has degraded too much *i.e.* not all zones can be reached by an available ambulance within r_2. In order to avoid infeasible solutions, we relax the covering constraints and strongly penalize their violation in the objective function. Also, to prevent a specific ambulance from being relocated too often, we include a penalty term for ambulances that have been relocated recently. The solution obtained indicates where the available ambulances should be placed. Then, which specific ambulances to redeploy are identified by minimizing the total traveled distance which consists in solving a min-cost max-flow problem. Finally, when a new emergency demand occurs, the nearest available ambulance is sent.

Transfer demands, on the other hand, are scheduled to form "routes" (*i.e.* a sequence of transfers to do). Routes are designed using a tabu search algorithm similar to the one proposed in Kergosien et al. (2011). It uses a lexicographic objective function which first minimizes the sum of transportation delays and crew overtimes then the sum of traveled distances, the motivation being that in practice the first criterion is often more important than the second (the latter serving only to avoid useless return trips). Since requests arrive dynamically, all routes are recomputed each time a new demand occurs or an existing demand is cancelled.

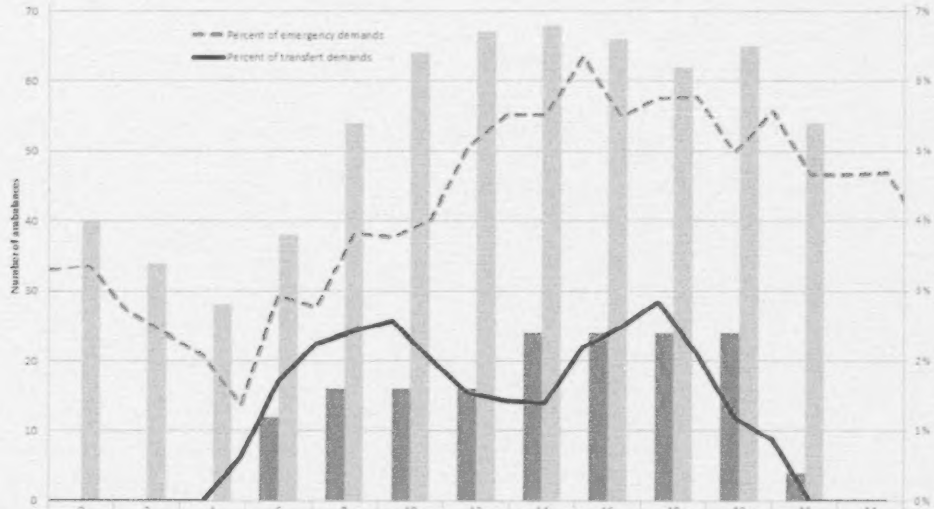
Decision procedures of the complete pooling fleet management: In the case of a single pooled fleet, all vehicles are considered as in the case of the emergency fleet described above. Whenever a transfer demand arrives, it is modeled as a "dummy" emergency demand that will appear at the starting time of the time window during which the transport has to begin. The nearest available ambulance will be assigned to it.

5.4 Results

The objective of the experiments presented here is first to verify and validate the simulation model by assessing how the performances obtained through the simulator correspond to the expected ones as well as through consistency analyses. The second objective of these experiments is to illustrate the flexibility and capability of the simulation tool, *i.e.* that the simulation tool is able to model several management strategies and measure their expected performance for a given context. In particular, this shows that the simulation model works efficiently and can be adapted to replicate the two fleet management strategies described previously. However, we want to stress that these experiments are in no way intended to be a thorough comparison of the alternate management strategies considered nor to compare or judge existing methods for solving the different optimization sub-problems present in the context of ambulance fleet management (i.e. dial-a-ride problem or ambulance relocation problem).

The experiments are structured as follows. For each management strategy (independent fleets and pooled fleet), we explored two cases having a total of 150 and 200 paramedical teams, respectively. By doing so we wish to assess if and how much system congestion (i.e. when reducing the amount of resources) translates into worse performance values. We assume that teams work each day, on 8-hour shifts. To set the number of paramedical teams on duty at each time of the day, we generated first a workload curve and then we assigned teams and vehicles to time slots in order to respect standard working constraints (i.e. maximum shift length and lunch breaks). Figure 7 shows the number of paramedic teams on duty depending on the time of day for the case where 200 ambulances were considered.

In order to verify and validate the simulation model, several performance measures were recorded over the simulation runs. We report hereafter those measures we deemed the most relevant with respect to the



	0	2	4	6	8	10	12	14	16	18	20	22	24
Ambulance for transfer demands	0	0	0	12	16	16	16	24	24	24	24	4	0
Ambulance for emergency demands	40	34	28	38	54	64	67	68	66	62	65	54	0

Figure 7: Number of ambulances during a day

objectives of the computational experiments. Results reported in Table 3 are based on 20 replications, each one composed of the same 7 consecutive simulated days. In order to remove the transient states corresponding to the first and last day of the horizon, the reported results are computed using only the 5 middle days. The time required to perform a simulation run (*i.e.* the 20 replications) is about 2 hours for the independent strategy and 20 minutes for the complete pooling strategy. It should be stressed, however, that the essential part of the overall computation time is due to the running times of the algorithms that replicate relocation and routing decisions. In particular, the tabu search algorithm used to take the routing decisions for transfer demands in the independent strategy requires considerable computing time which also explains why there is such a significant difference between the running times of the two fleet management strategies.

For each measure, the values reported are the average over the 20 replications and the standard deviation within parenthesis. As mentioned earlier, special attention was given to the use of Variance Reduction Techniques (Law and Kelton, 2000) like using *common random numbers* for the different replications. As a result, the standard deviations for performance measures are generally small even for only 20 replications.

Let us first analyze the results for the 200 ambulances case. In terms of service, both strategies handled an average 2593.4 urgent and 649.7 transfer requests. Both management strategies were able to respond to approximately 75 % of emergency demands within the 9 minutes threshold and around 90% within 11 minutes, the average response time for urgent requests being 439.7 and 433.6 seconds for the independent fleets and pooled fleet, respectively. We therefore observe very similar performances with a slight advantage to the pooled fleet. Unsurprisingly, these performances are significantly below the theoretical ones embedded in the deterministic relocation mathematical model (which was configured to ensure response to 90 % of the demands in less than 9 minutes and 100 % in less than 11 minutes). However, this gap between theoretical and empirical performances is easily explained by the strong variability of emergency demands which is not taken into consideration by the relocation model.

As for the transfer demands, the complete pooling strategy clearly outperforms the independent fleets. This was expected because the pooling strategy allows any ambulance to serve any demand thus increasing the system flexibility. In other words, the system can take advantage of any lull in emergency demand

	Fleet of 200 ambulances		Fleet of 150 ambulances	
	Fleet management strategies		Fleet management strategies	
Emergency demands	Independent	Complete pooling	Independent	Complete pooling
Number of emergency demands	2593.4 (49.3)	2593.4 (49.3)	2600.8 (40.4)	2600.8 (40.4)
Response time (R.T) in sec.	439.7 (4.1)	433.6 (4.0)	514.9 (13.9)	541.0 (15.2)
Percentage of demands such that R.T.$\leq$ 540 sec.	74.4 (0.81)	75.7 (0.78)	57.5 (1.5)	58.1 (2.0)
Percentage of demands such that R.T.$\leq$ 660 sec.	90.0 (0.62)	90.7 (0.68)	75.3 (1.7)	75.2 (1.8)
Transfer demands				
Number of transfer demands	649.7 (18.2)	649.7 (18.2)	653.9 (15.1)	653.9 (15.1)
Number of late transfer demands	69.7 (20.88)	1.6 (1.6)	265.2 (36.0)	52.9 (18.7)
Delays per late requests in sec.	3266.2 (1232.5)	310.4 (191.5)	6658.8 (830.1)	1221.6 (361.3)
Ambulances				
Number of diversions	1949.8 (87.1)	2333.2 (120.4)	2792.2 (80.2)	3250.4 (75.1)
Number of relocations performed	652.8 (30.2)	682.4 (32.5)	490.5 (19.1)	432.9 (23.9)
Percentage of time ambulances perform empty travels	20.4 (0.3)	23.4 (0.2)	23.2 (0.2)	25.0 (0.2)
Percentage of time ambulances are occupied	44.8 (0.9)	44.7 (0.9)	59.4 (0.9)	59.8 (1.0)
Number of overtimes for paramedics	614.9 (5.4)	599.6 (6.8)	481.1 (5.9)	461.0 (6.9)
Overtime for paramedics in sec.	3659.7 (101.3)	3649.3 (102.7)	4558.4 (184.9)	4829.2 (213.9)

Table 3: Results

arrivals to serve waiting transfer demands. In particular, the pooled fleet performed only 1.6 late transfers, on average, while the independent fleet incurred 69.7 for the same number of transfer requests. The average delay of late transfers were of around 5 minutes in the case of the pooled fleet, but of almost 55 minutes in the independent fleet case. We were surprised by the large values of the average and the standard deviation of delays produced by the independent fleet, but after a thorough look at the results, we realized that this was the result of some transfer demands not being able to be performed during the day and that had to be postponed to the next day (thus resulting in delays are larger than 28800s). Although this behaviour does not make much sense in practice (overtime could/should be used to complete all the demands), it confirms that the simulator reproduces with fidelity the management rules proposed by the user. It is also worth to mention that the similar performances produced by the two management strategies were also expected since the size of each fleet in the independent fleets case was "optimized", in other words, they were selected to fit the demand curve as described previously.

Regarding the efficiency of the fleet, ambulance occupation was around 45% for both strategies, the pooled fleet performing in average a few more diversions and relocations than the independent fleet (2332.2and 682.4 as opposed to 1949.8 and 652.8). These results were expected because (1) diversions might occur frequently in the complete pooling strategy due to the priority of emergency demands over are transfer demands, and (2) diversion is limited to the urgent fleet when they are managed separately. The increase in both the number of diversions and relocations directly leads to an increase in the percentages of time ambulances perform empty travels. This could be seen as the "price of flexibility".

The right part of Table 3 shows how performance deteriorates when the number of ambulances is reduced from 200 to 150, leading to a more *saturated* system with higher occupation rate (almost 60%). Average response times increase to 514.9 and 541.0 seconds while the percentage of demands responded within 9 minutes falls to 57.5% and 58.1%, for independent fleets and pooled fleet, respectively. Similarly, the number of late transfers as well as the average delay increase for both management strategies. In such saturated conditions, there is not much opportunity left to perform relocations, because ambulances are occupied too often. Nonetheless, more and more diversions are performed as a way to respond to the

arrival of urgent demands.

The analysis of these results, allow us to conclude that the proposed simulation model indeed succeeds in adequately representing a complex EMS context as well as two very different fleet management strategies such as what was considered in this study, and this under different conditions (number of ambulances). It therefore successfully illustrates the flexibility and usefulness of the proposed simulation model confirming thus the great potential of such analytic tools for investigating the performance of varied and eventually quite complex management strategies in diverse EMS contexts.

The results also seem to indicate that the use of a pooling strategy can be a very interesting management alternative in the context of a large EMS as the one modeled here.

6 Conclusion

EMS management generally involves many challenging decisions. This is mainly due to the highly random and dynamic nature of the system. To help EMS managers in their decision-making process, the development of a simulation-based analysis tool can be very useful. Indeed, simulation has been proven to be an effective analysis methodology to compare different scenarios while integrating stochastic and dynamic aspects. In this context, this paper presents a generic simulation model that can adequately represent all important operations related to the management of an EMS. One interesting aspect of this simulation model is that it explicitly considers the two possible types of tasks carried out by EMS: emergency demands and transfer demands. The proposed discrete-event simulation model is highly flexible thus allowing the analysis of several management strategies for both types of demands, regardless of the specific geographical characteristics of the geographical area being considered. The computational experiments presented to verify and validate the simulation model have also shown its relevance and capability to adequately represent the different aspects of the EMS context considered here. However, we want to emphasize that the simulation model presented has been designed in order to be easily adapted to handle diverse contexts. What we have presented in this paper is in fact a simulation analysis tool which can be used to compare several strategies related to decision problems faced by EMS. Among others, relocation and fleet management strategies can be investigated using this tool.

References

Aboueljinane L., E. Sahin, Z. Jemai 2013. "A review on simulation models applied to emergency medical service operations." *Computers & Industrial Engineering* 66: 734-750.

Altiok T., B. Melamed 2007. "Simulation Modeling and Analysis with ARENA." Academic Press.

Andersson T., S. Petersson, P. Värbrand 2007. "Decision support for efficient ambulance logistics." ITN Research Report LiTH-ITN-R-2005-1, Linkopings Universiteit.

Beaudry A., G. Laporte, T. Melo, S. Nickel 2009. "Dynamic transportation of patients in hospitals." *OR Spectrum* 32: 77-107.

Bélanger V., A. Ruiz, P. Soriano 2012. "Déploiement et redéploiement des véhicules ambulanciers dans la gestion des services préhospitaliers d'urgence." *INFOR* 50: 1-30.

Berlin G.N., J.C. Liebman 1974. "Mathematical analysis of emergency ambulance location." *Socio-Economic Planning Sciences* 8: 323-328.

Brotcorne L., G. Laporte, F. Semet 2003. "Ambulance location and relocation models." *European Journal of Operational Research* 147: 451-463.

Carpentier G. 2006. "La conception et la gestion d'un réseau de service ambulancier." Mémoire de maitrise, Université Laval.

Fujiwara O., T. Makjamroen, K.K. Gupta 1987. "Ambulance deployment analysis : A study case of Bangkok." *European Journal of Operational Research* 31: 9-18.

Gendreau M., G. Laporte et F. Semet 2006. "The maximal expected relocation problem for emergency vehicles." *Journal of the Operational Research Society* 57: 22-28.

Gendreau M., G. Laporte, F. Semet 2001. "A dynamic model and parallel tabu search heuristic for real-time ambulance relocation." *Location Science* 27: 1641-1653.

Gendreau M., G. Laporte, F. Semet 1997. "Solving an ambulance location model by tabu search." *Location Science* 5: 75-88.

Goldberg J.B. 2004. "Operations Research Models for the Deployment of Emergency Services Vehicles." *EMS Management Journal* 1: 20-39.

Goldberg J.B., R. Dietrich, J. M. Chen, M. G. Mitwasi 1990. "A simulation model for evaluating a set of emergency vehicle base locations : Development, validation and usage." *Socio-Economic Planning Sciences* 24: 124-141.

Hanne T., T. Melo, S. Nickel 2009. "Bringing Robustness to Patient Flow Management Through Optimized Patient Transports in Hospitals." *Interfaces* 39: 241-255.

Harewood S.I., S. Budge, E. Erkut 2002. "Emergency ambulance deployment in Barbados : a multi-objective approach." *Journal of the Operational Research Society* 53: 185-192.

Henderson S., A. Mason 2005. "Ambulance Service Planning: Simulation and Data Visualisation." *Operations Research and Health Care* 70: 77-102.

Hogan, K.,C.S. ReVelle 1986. "Concepts and applications of backup coverage". *Management Science* 34: 14341444

Ingolfsson A. 2013. "EMS Planning and Management." In *Operations Research and Health Care Policy*, Springer, New York, 105-128.

Ingolfsson A., E. Erkut, S. Budge 2003. "Simulation of single start station for Edmonton EMS." *Journal of the Operational Research Society* 54: 736-746.

Kergosien Y., Ch. Lenté, D. Piton, J.-C. Billaut 2011. "A tabu search heuristic for the dynamic transportation of patients between care units." *European Journal of Operational Research* 214: 442-452.

Kiechle G., K.F. Doerner, M. Gendreau, R.F. Hartl 2008. "Waiting Strategies for Regular and Emergency Patient Transportation." *Operations Research Proceedings* 6: 271-276.

Kleijnen J.P.C. 1995. "Verification and validation of simulation models." *European Journal of Operational Research* 82: 145-162.

Kelton W.D., A.M. Law 2000. "Simulation Modelling Analysis." Third edition, McGraw-Hill, New York.

Lewis P.A.W., G.S. Shedler 1979. "Simulation of Nonhomogeneous Poisson Process by Thinning." *Naval Research Logistics Quarterly* 26: 403-413.

Liu M..S., J.T. Lee 1988. "A simulation model of a hospital emergency call system using SLAMII." *Simulation* 51: 216-221.

Lubicz M., B. Mielczarek 1987. "Simulation modelling of emergency medical services." *European Journal of Operational Research* 29: 178-185.

Marianov V., C.S. ReVelle 1995. "Siting emergency services." In *Facility Location. A survey of Applications and Methods*, Z. Drezner, Ed. New York : Springer.

Mason A.J. 2013. "Simulation and Real-Time Optimised Relocation for Improving Ambulance Operations." In *Hanbook of Healthcare Operations Management: Methods and Applications*, B. T. Denton, Ed. New York : Springer.

Parragh S.N. 2011. "Introducing heterogeneous users and vehicles into models and algorithms for the dial-a-ride problem." *Research Part C: Emerging Technologies* 19: 912-930.

Rajagopalan H.K., C. Saydam, J. Xiao 2008. "A multiperiod set covering location model for dynamic redeployment of ambulances." *Computers & Operations Research* 35: 814-826.

Repede J.F., J.J. Bernardo 1994. "Developping and validating a decision support system for location emergency medical vehicles in Louisville." *European Journal of Operational Research* 75: 567-581.

ReVelle C.S., D. Bigman, D. Schilling, J. Cohon, R. Church 1989. "Review, extension and prediction in emergency services siting models." *European Journal of Operational Research* 40: 58-69.

Savas E.S. 1969. "Simulation and cost-effictiveness analysis of New York's emergency ambulance service." *Management Science* 15: 602-627.

Schmid V. 2012. "Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming." *European Journal of Operational Research* 219: 611-621.

Statistics Canada. 2012. Montreal, Quebec (Code 2466023) and Quebec (Code 24) (table). Census Profile. 2011 Census. Statistics Canada Catalogue no. 98-316-XWE. Ottawa. Released October 24, 2012. http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lang=E.

Swoveland C., D. Uyeno, I. Vertinsky, R. Vikson 1973. "A simulation model-based methodology for optimization of ambulance service policies." *Socio-Economic Planning Sciences* 7: 697-703.

Trudeau P., J.M. Rousseau, J. A. Ferland, J. Choquette 1989. "An operations research approach for the planning and operation of an ambulance service." *INFOR* 27: 95-113.

Emergency Medical Services Systems Act of 1973. (P.L. 92-154). 93rd Congress.

Urgences-Santé, rapports annuels, http://www.urgences-sante.qc.ca/.

Zhen, L., K. Wang, H. Hu, D. Chang 2014. "A simulation optimization framework for ambulance deployment and relocation problems." *Computers & Industrial Engineering*, 72: 12-23.